# BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

## IMAGES ARE BEST AVAILABLE COPY.

**As rescanning documents *will not* correct images, please do not report the images to the Image Problem Mailbox.**

# Subjective Evaluation of MPEG-4 Video Codec Proposals: Methodological Approach and Test Procedures.

**T. Alpert[1], V. Baroncini[2], D. Choi[3], L. Contin[4], R. Koenen[5], F. Pereira[6], H. Peterson[7]**

[1] Centre Commun díEtudes de Télédiffusion et Télécommunications 4, rue du Clos Courtel - BP 59, Cesson Sévigné Cedex, FRANCE

[2] Fondazione Ugo Bordoni, Viale Europa 190, 00149 Roma, ITALY

[3]Hughes Aircraft Co., 7200 Hughes Terrace, 90045 Los Angeles, CA, U.S.A.

[4]Centro Studi E Laboratori Telecomunicazioni, V. Reiss Romoli, 274, 10148 Torino, ITALY

[5]KPN Research, PO Box 421, 2260AK Leidschendam, THE NETHERLANDS

[6]Instituto Superior Técnico, Av. Rovisco Pais, 1096 Lisboa Codex, PORTUGAL

[7]David Sarnoff Research Center, 201 Washington Road, Princeton, NJ 08540-6449, U.S.A.

# Abstract

A new audio-visual coding standard, MPEG-4, is currently under development. MPEG-4 will address not only compression, but also completely new audio-video coding functionalities related to content-based interactivity and universal access. As part of the MPEG-4 standardization process, in November, 1995 assessments were performed on technologies proposed for incorporation in the standard. These assessments included formal subjective tests, as well as expert panel evaluations.

This paper describes the MPEG-4 video formal subjective tests. Since MPEG-4 addresses new coding functionalities, and also operates at bitrates lower than ever subjectively tested before on a large scale, standard ITU test methods were not directly applicable. These methods had to be adapted, and even new test methods devised, for the MPEG-4 video subjective tests. We describe here the test methods used in the MPEG-4 video subjective tests, how the tests were carried out, and how the test results were interpreted. We also evaluate the successes and shortcomings of the MPEG-4 video subjective tests, and suggest possible improvements for future tests. The MPEG-4 video subjective tests were successful, providing the MPEG community with critical information to guide in the selection of technologies for inclusion in the video part of the MPEG-4 standard.

# 1. Introduction

The primary purpose of audio-visual coding standards has historically been compression, that is, efficient representation of audio-visual information. However, new developments taking place in the production and handling of audio-visual data have given rise to new demands on coding standards. Audio-visual data production methods are becoming more sophisticated, incorporating more synthetic material. Audio-visual information consumption is increasingly interactive. New types of networks are carrying audio-visual information, from mobile narrowband to broadband. And hardware and software technologies continue to progress. To address these new developments, the MPEG committee has defined the following set of functionalities that the MPEG-4 standard will target:

1. Content-based multimedia data access tools

2. Content-based manipulation and bitstream editing

3. Hybrid natural and synthetic data coding

4. Improved temporal random access

5. Improved coding efficiency

6. Coding of multiple concurrent data streams

7. Robustness in error-prone environments

8. Content-based scalability

In July of 1995 the MPEG committee issued a call for proposals to the international technical community for technology for inclusion in the MPEG-4 standard. Proposals for video coding schemes were received from thirty-two different technical teams from around the world. These proposals were evaluated via a series of formal subjective tests, to assess the merit of the proposals relative to one another, and also relative to existing standards.

Due to time limitations and the large volume of proposals anticipated, the scope of the subjective tests was concentrated on three "representative functionalities," selected to represent the eight functionalities of MPEG-4: improved coding efficiency, robustness in error-prone environments, and content-based scalability. The subjective tests were conducted during a meeting of the MPEG Ad Hoc Group on MPEG-4 Video Testing Logistics, held from October 30th to November 3rd, 1995, at the Hughes Electronics Corporation in Los Angeles, California. More than 70 video and test experts from more than 20 countries participated in the proposal evaluation process.

In addition to the formal subjective tests mentioned above, for some content-based scalability proposals, evaluations were also performed by panels of video experts. Expert panels also evaluated algorithm proposals not specifically supporting any of the three representative functionalities, and proposals for coding tools (that is, components of coding algorithms, for example, motion compensation or contour coding). This paper only addresses the methodology and the test procedures used for the MPEG-4 video formal subjective tests. Details about the expert panel evaluations are contained in [1] and [2].

The goals of MPEG-4 presented new challenges in the design of effective subjective tests. Most of the MPEG-4 functionalities are new to audio-visual coding, and there was little prior experience in subjectively testing codec performance with regards to these functionalities. Moreover, the relatively low target bitrates for MPEG-4 (from 10 to 1024 kbit/s) required tests to assess video with quality significantly below that of broadcast video. However, traditional subjective testing methods (such as those specified in [3]) were designed mainly to evaluate broadcast quality video material. These factors dictated that traditional subjective testing methods be tailored for use in testing the MPEG-4 proposals.

This paper describes the methodological approach, organization, test procedures, and data analysis used in the MPEG-4 subjective tests. In Section 2 we describe basic subjective test methods, and how these methods were adapted for use in the MPEG-4 tests. The lab set-up and test materials used for the tests are described in Sections 3 and 4, and Section 5 describes the proposal specifications. Section 6 gives general information on the organization of the tests, and Section 7 details the procedures used for each specific subjective test performed on the MPEG-4 video proposals. The statistical analysis of the subjective test data is described in Section 8. Finally, in Section 9 we evaluate the effectiveness of the MPEG-4 test methods, and make suggestions for how they might be improved upon for future tests. Further details on the subjective tests, and the results of the statistical analysis of the test data are found in [4].

# 2. Test Methods

As stated above, in order for the MPEG-4 subjective tests to be fair and effective, traditional subjective test methods had to be adapted. In designing the tests, several factors had to be taken into account: 1) Completely new coding functionalities were to be evaluated, 2) Due to the low target bitrates for MPEG-4, video with quality significantly below broadcast quality was to be evaluated, and 3) Due to the new kinds of object-based algorithms considered by MPEG-4 and the channel error conditions that MPEG-4 takes into account, a possibly large number of different types of impairments were to be evaluated.

In order to account for these factors in designing the subjective tests, several measures were taken: 1) For some of the new MPEG-4 functionalities, besides the usual evaluation of the global video quality, new tasks were defined for the test subjects. For example, in one test, test subjects were instructed to evaluate the quality of a single object within a scene instead of the quality of the whole scene. The new tasks are described completely in Section 7. 2) The MPEG-4 test bitrate range was divided in three main sub-ranges and, generally speaking, different test methods were applied for the different sub-ranges. In addition, modifications to the standard test methods were introduced in order to optimize them to the case of clearly detectable impairments. All the MPEG-4 test methods are described below. 3) The effects of the wide range of different impairments were minimized by taking particular care in the test subjects' training. For example, during training for each test, subjects were shown sample video sequences having qualities over the full range expected in that test. Test subject training is described fully in Section 6.

Four test methods were used in the MPEG-4 video subjective tests: double stimulus continuous quality scale, double stimulus impairment scale, double stimulus binary vote, and single stimulus. These methods are described in detail below.

### 2.1 Double Stimulus Continuous Quality Scale (DSCQS) Method

The presentation sequence for a DSCQS test trial is illustrated in Figure 1. In the DSCQS method, each trial consists of a pair of stimuli: one stimulus is the reference, and the other is the test. The test stimulus is usually the reference after undergoing some type of processing. The two stimuli are each presented twice in a trial, in alternating fashion, with the order of the two randomly chosen for each trial. To aid the test subjects in staying on track in their assessments, audio cues are used to indicate when a trial begins, when a new stimulus begins, when to vote, and what the current trial number is in the sequence of trials making up a test session. These audio cues are also illustrated in Figure 1.

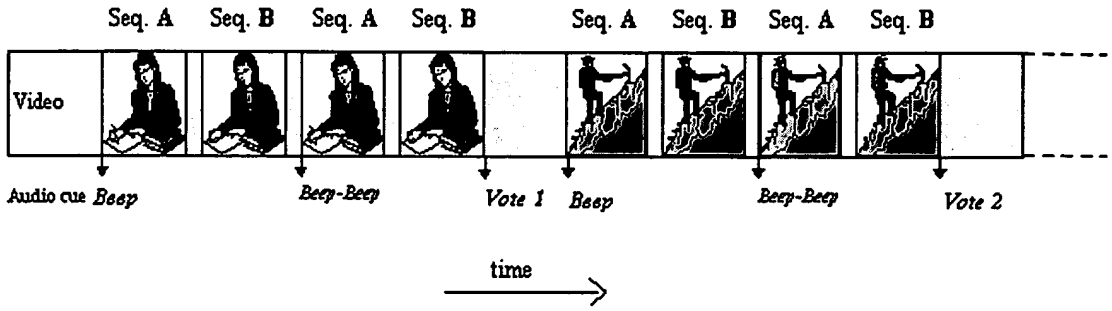
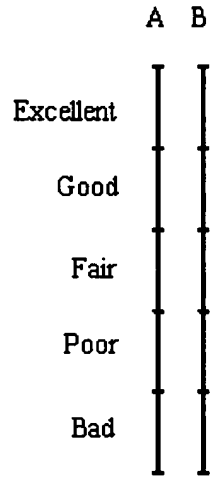



*Figure 1. Presentation sequence for the DSCQS test method.*




Test subjects are not informed of the ordering of the test and reference stimuli, and they rate each stimulus by marking a continuous quality scale. Thus, two ratings are made for each trial in the DSCQS method: one for the reference and the other for the test condition. An example of the rating scales for one DSCQS trial is given in Figure 2. Occasionally in a DSCQS test, both stimuli presented in a trial are the reference stimulus. Such trials are used to detect erratic test subject behavior.

*Figure 2. Rating scales for a trial with the DSCQS method.*

The DSCQS method is typically applied for evaluations where the quality difference between the test and reference sequences is not too large. Thus, in the MPEG4 tests, this method was applied to the highest bitrates (512 kbit/s, 1024 kbit/s, and sometimes 320 kbit/s, as specified in Section 7).

2.2 Double Stimulus Impairment Scale (DSIS) Method

The presentation sequence for a DSIS test trial (including audio cues) is illustrated in Figure 3. As in the DSCQS method, each trial consists of a pair of stimuli: the reference and the test. However, in the DSIS method, the two stimuli are always presented in the same order: the reference is always first, followed by the test.

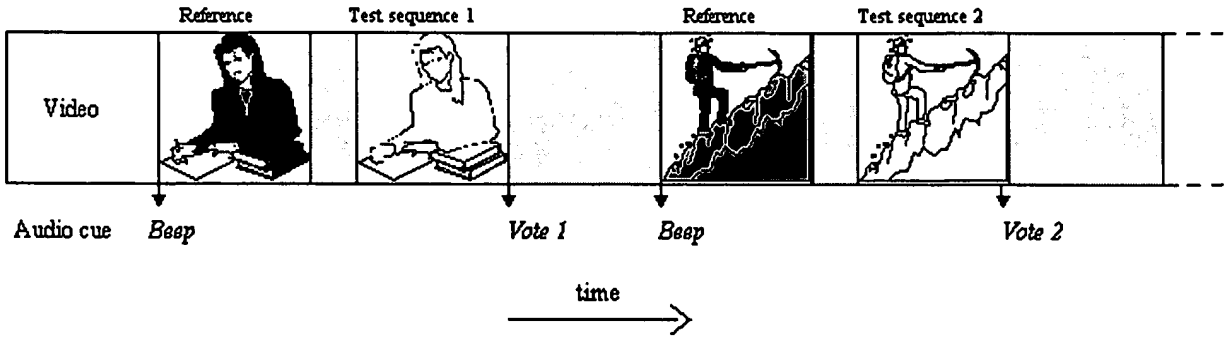


*Figure 3 - Presentation sequence for the DSIS method*

In the DSIS method, test subjects compare the two stimuli in a trial and rate the impairment of the test stimulus with respect to the reference, using a five-level degradation scale. Thus, only one vote is made for each DSIS trial. The rating scale used for one DSIS trial in the MPEG-4 tests is shown in Figure 4.

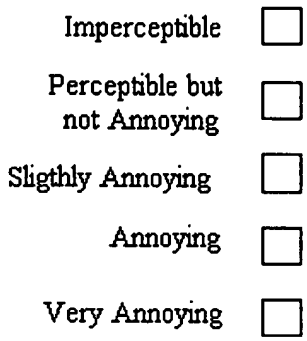



*Figure 4. Rating scale for the DSIS test method.*

The DSIS method is typically applied for evaluating the annoyance of video impairments; so it is primarily suited for evaluating the performance of systems that introduce clearly visible impairments. For this reason it was applied for evaluating the performance of proposals at low and medium bit rates (112 kbit/s, and sometimes 48 kbit/s and 320 kbit/s, as specified in Section 7).

Since the evaluation in the DSIS method is made with respect to a reference, it is important that the quality level of the test stimulus and the corresponding reference are not too different, or the usefulness of the reference could be greatly reduced. In the MPEG-4 tests, the original sequences were generally used as references. However, in order to maintain similar levels of quality in the test and reference stimuli, in many cases the reference sequences were displayed in a downsampled format, such as CIF. Details about the reference formats used are given in Section 7.

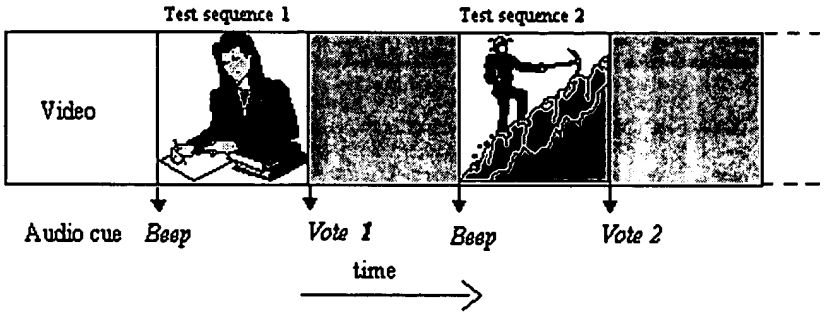
At the lowest MPEG-4 bitrates, downsampling sometimes did not bring the qualities of the original and decoded sequences sufficiently close. In such cases, since no method to produce appropriate alternative reference sequences was readily available, the DSIS method was not used.

2.3 Double Stimulus Binary Vote (DSBV) Method

The DSBV method is a new, non-standard test method, specifically designed by MPEG test experts to evaluate the performance of a codec following a very long burst of bit errors causing loss of codec synchronization. Very bad decoded video quality was expected under these conditions. Therefore, asking test subjects to evaluate video quality or video impairment was expected to be ineffective. Instead, test subjects were asked to evaluate whether the codecs under test recovered from the error burst by the end of the test sequence.

A presentation sequence similar to that of the DSIS method was used in the DSBV tests (see Figure 3). A video sequence decoded under error-free conditions was used as the reference, and the same sequence decoded in the presence of bit errors was used as the test. Test subjects compared the error- corrupted decoded sequence to the error-free decoded sequence, and voted "yes" or "no," to reflect whether or not the corrupted sequence had recovered from the error burst by the end of the sequence.

2.4 Single Stimulus (SS) Method




The presentation sequence for a SS test trial (including audio cues) is illustrated in Figure 5. In the SS method, only one stimulus is presented in each trial. The test subject rates each stimulus, typically using a five-level quality scale (i.e. Excellent, Good, Fair, Poor, Bad).

*Figure 5. Presentation sequence for the SS test method.*

The SS method is appropriate when references are not available. As indicated in Section 2.2, at the lowest tested bitrates appropriate reference sequences were not readily available. Thus, the SS test method was applied for evaluating the performance of proposals at very low bitrates (10 kbit/s, 24 kbit/s, and sometimes 48 kbit/s, as specified in Section 7).

A concern when testing low quality video sequences is compression of test subjects' votes toward the bottom of the rating scale, making quality discriminations between codecs more difficult. Since a larger number of rating levels can increase the discriminative power of the scale, instead of the usual five-level scale, MPEG-4 used an eleven-level scale, illustrated in Figure 6. As mentioned above, during test subject training prior to each MPEG-4 test, test subjects were shown sample video sequences having qualities over the full range expected in each test. This training served to further reduce the risk of vote compression in the SS tests.

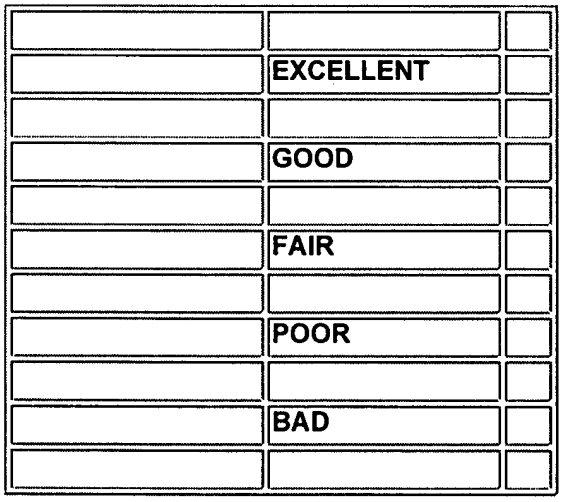



*Figure 6 - Rating scales used in MPEG-4 for the SS method*

Since explicit references are not used in SS methods, context dependency (the effect of previously seen trials on the rating given to the current trial) is stronger than for test methods which use an explicit reference. To compensate for this effect, each SS test was performed twice, with two different trial presentation orders.

# 3. Test Laboratory Set-up

The MPEG-4 tests were conducted in a specially built lab at the Hughes Electronics Corporation in Los Angeles, California, over three days. Five identically configured test stations were used simultaneously to carry out the tests. Figure 7 shows the general layout of the five test stations in the lab.

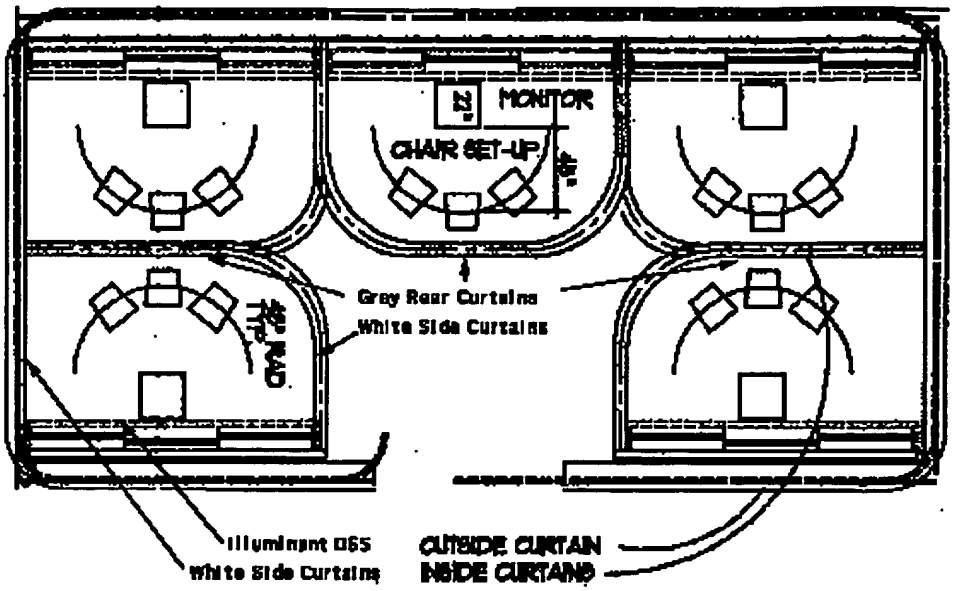



*Figure 7: Layout of the test stations*

The test environment for the MPEG-4 tests was in accordance with [3]. The viewing conditions set forth by this standard include:

- the monitor brightness and contrast (set using the PLUGE [5] technique)
- the ratio of the inactive monitor luminance to the peak monitor luminance
- the ratio, in a completely dark room, of the monitor luminance when displaying black to the peak monitor luminance
- the ratio of the luminance of the background behind the monitor to the peak monitor luminance
- the chromaticity of the background behind the monitor
- the room illumination
- the maximum test subject observation angle relative to monitor screen normal.

At each test station, three test subjects sat in an arc arrangement, with all test subjects less than 30 degrees off-axis from the monitor screen normal. Each test subject's viewing distance was four picture heights (45 inches). The equipment at each test station consisted of a SONY BVM-1911 RGB (19 inch diagonal) monitor, and a loudspeaker. Figure 8 shows a picture of a test station.

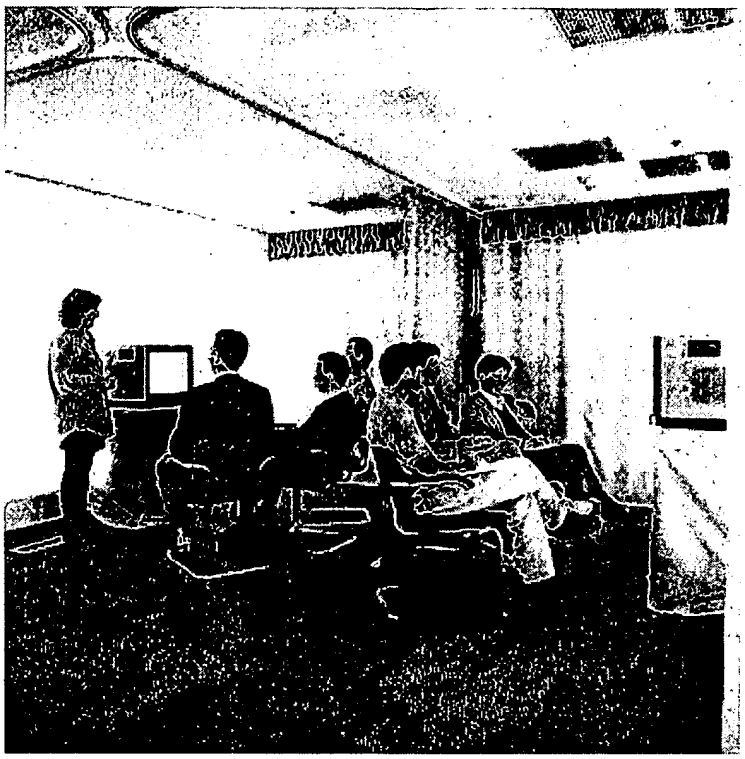


*Figure 8: One test station.*

The test material for the subjective tests was professionally edited onto D-1 tapes, according to pre-specified test scripts. During each test session, the appropriate tape was played on a D-1 VTR in 525/60 Hz format, and the video output was distributed to the five test station monitors using an Extron Video Distribution Amp. Identical audio, recorded on each of the four audio channels of the D-1 tapes, provided the audio cues to four of the five test stations during the tests. Audio for the fifth test station was provided by recording onto one of the other output ports on the D-1 VTR, for example the monitor out or the audio cue track. A monitor and loudspeaker in the D1 VTR room provided the capability to monitor the video and audio distributed to each test station. Figure 9 shows the equipment layout for the tests. On the left is the equipment in the D-1 VTR room, and on the right is the equipment at one test station.

For calibration purposes, each D-1 tape contained 100% color bars and a PLUGE pattern, generated as specified in [5]. The test monitors were calibrated in accordance with the ITU-R Recommendation [5].

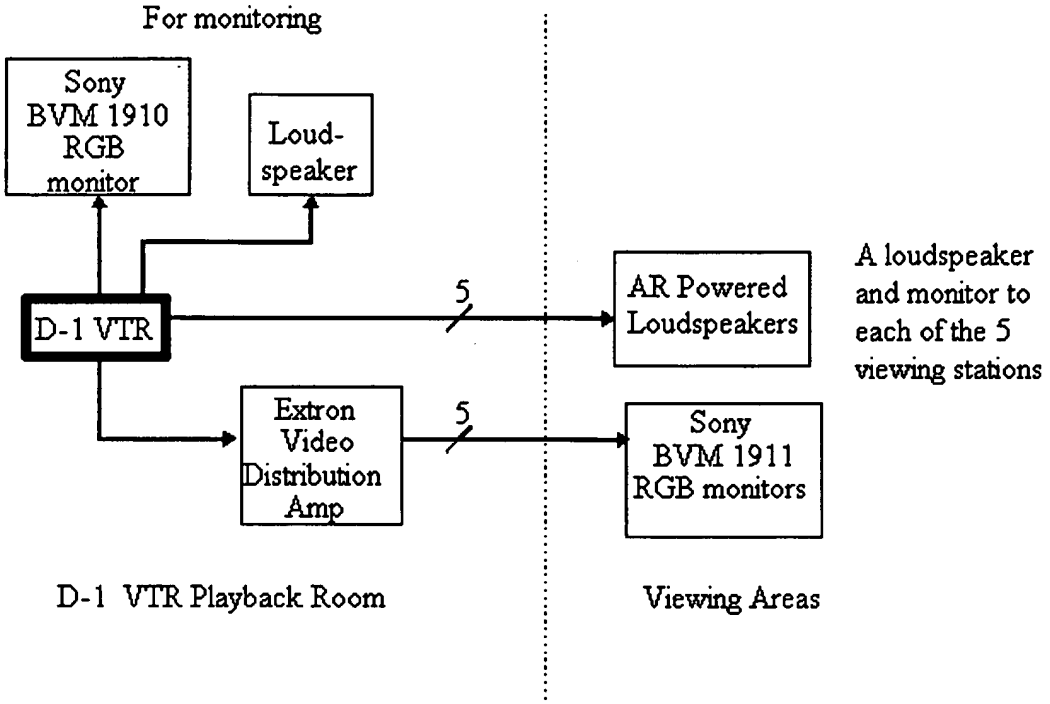



*Figure 9: Test equipment configuration*

# 4. Test Materials

For the MPEG-4 video tests, a library of test material was established and made available to the technical community (proposers and other interested parties) via a combination of exabyte tapes and ftp sites. The library contained original video sequences in ITU-R 601 format [6], captured at either 60 or 50 Hz. The original sequences in the library were divided into five classes, according to characteristics of their spatio-temporal content. Table 1 lists the five sequence classes, the sequences in each class, and the frame rate at which each sequence was captured.

| Sequence class | Resolution | Content complexity | Test sequences (sampling rates) |
|---|---|---|---|
| A | ITU-R 601 | Low spatial detail and low amount of motion | Mother & daughter (60 Hz),<br>Akiyo (60 Hz),<br>Hall monitor (60 Hz),<br>Container ship (60 Hz),<br>Sean (60 Hz) |
| B | ITU-R 601 | Medium spatial detail and low amount of motion or vice versa | Foreman (50 Hz),<br>News (60 Hz),<br>Silent voice (50 Hz),<br>Coast guard (60 Hz) |
| C | ITU-R 601 | High spatial detail and medium amount of motion or vice versa | Table tennis (60 Hz),<br>Stefan (60 Hz),<br>Mobile & calendar (60 Hz),<br>Fun fair (left view of stereoscopic sequence, 50 Hz) |
| D | ITU-R 2*601 | Stereoscopic | Tunnel (50 Hz),<br>Fun fair (50 Hz) |
| E | ITU-R 601 | Hybrid natural and synthetic | Children (60 Hz),<br>Bream (60 Hz),<br>Weather (60 Hz),<br>Destruction (60 Hz) |

Table 1: MPEG-4 video test sequence library

Though some of the sequences listed in Table 1 are longer than 300 frames, for the MPEG-4 tests, only the first 300 frames of each sequence were used. During the tests, regardless of the original frame rate, all test sequences were displayed at 60 Hz field rate, yielding test sequences 10 s in length (since each test sequence was 300 frames). This compromise was made due to concerns that frame rate conversion could introduce degradation into the test sequences, biasing the test results. The distortion induced by displaying a 50 Hz sequence at 60 Hz was considered less objectionable than that induced by frame rate conversion.

In addition to the original video sequences described above, two other types of test materials were distributed. The first type of additional test material provided was segmented original sequences. Since many of the MPEG-4 functionalities are content-based, coding often requires segmentation of a sequence into meaningful objects. For the purposes of these tests, however, image segmentation was considered to be a pre-processing step, and not to be evaluated. Consequently, it was desired that the results of the subjective tests not be influenced by differing segmentation techniques among the proposals. Therefore, for each of the class A, B and C original video sequences used in the content-based scalability tests, a segmented version of each frame in the sequence was provided to the proposers for

use in their coding. The segmentations had a maximum of 256 segments per frame, and the segments had semantic meaning, that is, they corresponded to meaningful objects in the video scene. Proposers were encouraged to use the provided sequence segmentations; however, they were permitted to use other segmentations, so long as the segmentation met the requirements described above, and that it was clearly described in the proposal technical description.

The second type of additional test material provided to proposers was alpha plane sequences for class E sequences. Class E sequences are composed of natural and synthetic content, understood as a composition of multiple layers. (In MPEG-4 terminology, such a layer is referred to as a Video Object Plane (VOP) [7].) For a class E sequence, each alpha plane sequence gives a value (between 0 and 255) for each pixel, describing the blending contribution of the corresponding VOP for that pixel.

# 5. Proposal Specifications

An algorithm proposal submitted for an MPEG-4 subjective test consisted of four required elements:

1. A D-1 tape containing all required decoded video sequences for the test addressed, in 60 Hz ITU-R 601 format, with resolution as specified for that test. (Section 7 gives more specific details on the contents of this tape.)
2. An exabyte tape containing the executable decoder which implements the proposer's algorithm, and containing an encoded bitstream corresponding to each of the decoded sequences on the proposer's D-1 tape. The executable decoder must be capable of decoding the encoded bitstreams to produce the decoded video sequences on the proposer's D-1 tape.
3. A complete technical description of the proposal, including all elements necessary for a full understanding of the proposal's achievements, as well as the methods that allow it to reach the performance demonstrated on the proposer's D-1 tape.
4. The submission check list found in [8], completed with the requested information about the proposal and bitstream statistics.

The MPEG-4 test specifications placed very few restrictions on the processing and coding methods used by proposers. For example, proposers could used any desired techniques for pre-processing and post-processing, and downsampling and upsampling, as well as any spatial and temporal coding resolutions, provided that these were fully described in the proposal technical description. However, since it was expected that many proposals would include downsampling, suggested downsampling and upsampling filters with known good performance were provided [8]. This helped to make the test results easier to interpret, and also helped to ease the burden of work on the proposers. Proposers were free to use alternative downsampling and upsampling filters if they preferred, provided that a precise description was included in the proposal technical description.

# 6. Test Execution

6.1 Test Session Description

The MPEG-4 video tests were conducted via a series of test sessions, each lasting no longer than 30 minutes. When, due to a large number of proposals, a test required longer than 30 minutes to complete, the test was split into a series of sessions, each less than 30 minutes in length. (The longest test, compression of class A sequences, required six test sessions to complete.) A test session generally began with a training phase, followed by one or more testing phases, each composed of a stabilization period followed by the actual test.

During the training phase, typically lasting approximately five minutes, the test subjects first read written instructions describing the test method. (The texts for all the test instructions are found in the annex to [4].) The test instructions were also reviewed orally, and test subjects could ask questions. In order to preserve consistency between test sessions, care was taken when answering questions to only clarify the information in the written instructions, and to not inject any additional information into the instructions. To complete the test subject training, a brief "practice" test session (consisting of several test trials) was conducted. This served to familiarize the test subjects with the trial presentation sequence, the audio cues, the score sheets for recording their ratings, the assessment procedure, and the quality range of the video to be expected during the test. The test sequences and coding algorithms used during the practice session were different from those used in the actual test.

The testing phase of each test session began with five stabilization trials. The actual test trials directly followed the stabilization trials, without interruption. Test subjects were not informed of the presence of the stabilization trials, and the ratings made during the stabilization trials were not included in the test data. The stabilization trials used actual test sequences from the current test session, spanning the full quality range for that session. The purpose of the stabilization

period was to "stabilize" the subjects' judgments by exposing them to test sequences over the range of qualities which would be present in the test session. Following the stabilization trials, the actual test trials were presented in a pseudo-random order.

6.2 Anchor Sequences

The results of subjective quality assessments often depend not only on the actual video quality, but also on other factors such as the quality range of the test material, the experience and expectations of the test subjects, etc. In order to control these effects, whenever possible, the test sessions included appropriate "anchor" test sequences. These anchors were presented at random times in a test session. Test subjects were not informed of which trials were anchors, and the anchors were evaluated exactly as the proposers' test sequences. Anchors had several purposes in the tests. First, they served to expand and/or complete the range of qualities present in a test session. Second, they provided a means for comparing test subject behavior between test sessions for multi-session tests.

Anchors were decoded sequences generated using the best available standard codec operating at the same bitrates as the proposers' codecs, and were displayed using the same format as specified for the proposers' decoded sequences. By using existing coding standards, anchors served the additional purpose of providing a means for comparing the proposals to existing standards. Section 7 specifies the anchor used for each MPEG-4 test. H.263 anchors were generated using advanced coding options, with PB-frames switched on and off according to an adaptive scheme [9]. MPEG-1 anchors were generated according to [10].

6.3 Test Subjects and Scheduling

As described in Section 3, fifteen test subjects participated at once in a test session. In order to allow for rest time for the test subjects, but to also have tests running continuously, three teams of fifteen test subjects were used, for a total of 45 test subjects taking part in the MPEG-4 video tests. All test subjects were video coding experts, screened for normal color vision and normal or corrected-to-normal visual acuity. The large majority of test subjects were males between 20 and 40 years of age. The maximum active test time per day for a test team was two hours.

For tests split over several sessions, the same test team participated in all of the sessions making up that test. In these cases, or any other time a particular test method was used over multiple test sessions for the same test team, the training phase described above was only conducted for the first test session using that test method. For subsequent sessions using that method for those test subjects, the subjects were only briefly reminded of the test method and given an opportunity to ask questions.

# 7. Test Specifications

As was mentioned in the introduction, the subjective tests concentrated on three "representative" functionalities: improved coding efficiency, robustness in error-prone environments, and content-based scalability. Each of these tests is specified in detail below.

7.1 Improved Coding Efficiency

The first MPEG-4 functionality tested was "improved coding efficiency." According to [11], MPEG-4 shall provide "subjectively better audio-visual quality at comparable bitrates, compared to existing or emerging standards".

Four separate formal subjective tests were used to evaluate compression efficiency of the proposals. Table 2 gives, for each test, the test sequence class and test sequences, the test bitrates, the test method, the decoded sequence display format, the reference sequence format, the coding technique used for the anchor sequence, the number of proposals received to the test, and the number of test sessions used for the test. A proposer could address any subset of the tests listed in Table 2. However, for each test addressed, a proposer was required to encode all the test sequences at all the bitrates for that test (one row in Table 2). Due to time limitations, compression efficiency proposals addressing class D sequences (stereoscopic) were not evaluated with formal subjective tests, but instead were evaluated by panels of experts. For tests using a double stimulus method, the reference sequence was the original uncoded test sequence.

| Sequence class | Source test material | Bitrates (kbit/s) | Test method | Display format | Reference format | Anchor coding technique | Num.of proposals/ Num.of sessions |
|---|---|---|---|---|---|---|---|
| A: | Mother & daughter, Akiyo, Hall monitor, Container ship | 10, 24, 48 | SS | CIF (1) | N/A | H.263 | 17/5 |
| B: | Silent Voice, Foreman, News, Coast guard | 24, 48, 112 | SS | CIF (1) | N/A | H.263 | 13/4 |
| C: | Fun fair, Table tennis, Mobile & calendar, Stefan | 320, 512, 1024 | DSCQS | ITU-R 601 | ITU-R 601 | MPEG-1 | 6/4 |
| E: | Weather, Children, Bream, Destruction | 48, 112, 320 | DSIS | CIF (1) | CIF | H.263 (48&112kbit/s), MPEG-1 (320 kbit/s) | 4/2 |

(1) displayed in a window centered in the 60 Hz ITU-R 601 format.

Table 2 - Compression efficiency tests

7.2 Robustness in Error-prone Environments

The second MPEG-4 functionality tested was "robustness in error-prone environments." According to [11], MPEG-4 shall provide "an error robustness capability to allow access to applications over a variety of wireless and wired networks and storage media".

The error robustness tests assessed a codec's operation in error-prone environments by subjecting the proposers' coded bitstreams to channel bit error conditions representative of those in a variety of networks. Two classes of formal subjective tests of were used to evaluate error robustness of the proposed codecs: error resilience and error recovery. The error resilience subjective tests assessed the overall quality of video decoded in the presence of several different types of bit errors. The error recovery subjective tests assessed whether a codec was able to recover from severe bit errors in the encoded bitstream.

For the error resilience tests, three error conditions were used, typical of residual errors in a bitstream after network error control:

1. Random bit error rate (BER), in the interval [1.5 s, 10 s]
2. Three bursts of errors: random burst length 16 to 24 ms, 50% BER within burst, random burst separation greater than 2 s, in the interval [1.5 s, 8 s]

3. Error conditions (1) and (2) superimposed.

A proposer to an error resilience test provided three decoded versions of each test sequence, one version decoded under each of the above three error conditions. Subjective tests evaluated the global quality of these sequences.

For the error recovery tests, one error condition was used, a burst error followed by an error-free condition: One error burst: random burst length 1 to 2 s, 50% BER within burst, random burst start in the interval [1.5 s, 3 s]. The error recovery tests used the DSBV test method described in Section 2. This required that proposers supply two decoded versions of each test sequence: one version decoded in the presence of bit errors, and another version decoded under the error-free condition.

Table 3 gives, for each error robustness test, the test class (resilience or recovery), the test sequence class and test sequences, the test bit rate, the test method, the decoded sequence display format, the reference sequence format, the error condition(s), the number of proposals received to the test, and the number of test sessions used for the test. A proposer could address any subset of the tests listed in Table 3. However, for each test addressed, the proposer was required to encode all the test sequences under all the error conditions for that test (one row in Table 3). Anchors were not used in the error robustness tests, since no coding standard including error control techniques was readily available.

| Sequence class | Source test material | Functionality | Bitrates (kbit/s) | Test method | Display format | Reference format | Error conditions | Num.of proposals/ Num.of sessions |
|---|---|---|---|---|---|---|---|---|
| A: | Mother & Daughter, Akiyo, Hall monitor, Container ship | error resilience | 24 | SS | CIF (1) | N/A | random, burst, random+burst | 9/3 |
| B: | Silent Voice, Foreman, News, Coast guard | error resilience | 48 | SS | CIF (1) | N/A | random, burst, random+burst | 7/2 |
| A: | Mother & Daughter, Akiyo, Hall monitor, Container ship | error recovery | 24 | DSBV | CIF(1) | coded without channel errors | long burst | 7/1 |
| B: | Silent Voice, Foreman, News, Coast guard | error recovery | 48 | DSBV | CIF (1) | coded without channel errors | long burst | 6/1 |

(1) displayed in a window centered in the 60 Hz ITU-R 601 format.

Table 3 - Error robustness tests.

Error patterns for the error robustness tests were produced using the software in Annex E of [8], and then distributed to proposers. Proposers were prohibited from making any adjustments to their encoded

bitstreams or decoder simulations after subjecting their bitstreams to the error patterns. The error conditions used in the error robustness tests provided an initial error-free period, to allow transmission of an initial frame, and to allow the decoder to stabilize into a steady state.

As in all the subjective tests, the error robustness test sequences were 10 seconds long. However, for all the error robustness subjective tests, only the final 9 seconds of each 10 second decoded test sequence were used. This was done to exclude from the subjective evaluations the initial "start-up" performance of the proposed codecs.

Proposers had a great deal of freedom in the error control techniques used in their codecs. Techniques such as forward error correction, error detection, error containment, error concealment, and resynchronization were all permitted. Only two restrictions were imposed on codecs proposed to the error robustness subjective tests: (1) The decoder could only use information from the encoded bitstream; no out-of-band information could be used. (2) No feedback was permitted from the decoder to the encoder. Proposals not meeting these restrictions were not included in the formal subjective tests, but instead were evaluated by panels of experts.

Also, the following (non-mandatory) guidelines were suggested for proposed codecs [8]:

1. Initial codec delay: 1.0 s maximum
2. Instantaneous codec delay: 500 ms maximum, excluding first picture
3. Average codec delay: 250 ms maximum

Proposers of codecs not adhering to these guidelines were required to specify the above parameters in their proposal technical description.

7.3 Content-based Scalability

The third MPEG-4 functionality tested was "content-based scalability." According to [11], MPEG-4 shall provide "scalability with a fine granularity in content, quality (spatial resolution, temporal resolution), and complexity".

Formal subjective tests were used to evaluate two types of content-based scalability: object-based scalability, and quality scalability. Object-based scalability is the ability of a codec to vary the number of objects from a video sequence that are simultaneously decoded. With object-based scalability, encoding is performed in such a way that any one object in the video scene can be decoded independently of all other objects. Quality scalability involves the use of a scalable bitstream to represent objects in the video scene. With quality scalability, the bitstream is constructed so as to allow different subsets of the bits representing an object to be decoded, producing versions of the object with differing qualities. Two methods for accomplishing quality scalability are scaling the spatial resolution or scaling the temporal resolution of an object. These were the only types of quality scalability addressed in the subjective tests.

Proposals to the object-based scalability tests were tested using both formal subjective tests and expert panel evaluations. Subjective tests were used to assess the overall quality of a proposal's decoded video. Expert panel evaluations were used to assess a proposal's ability to decode a subset of video objects using a subset of the bitstream, as well as a proposal's ability to build a video scene object-by-object.

Proposals for quality scalability were evaluated with formal subjective tests at two different bitrates, representing two layers of scalability: a low bitrate "base layer," and a higher bitrate "base+enhancement layer." Different test approaches were used to assess the quality of the two layers. For the base layer evaluations, test subjects simply evaluated the overall quality of the low bitrate decoded video. For the base+enhancement layer evaluations, test subjects assessed only the quality of certain designated object(s) in the higher bitrate decoded video. The designated objects(s) in each test sequence were clearly identified to the test subjects before a base+enhancement layer test session began, and test subjects were instructed to confine their attention to only those object(s) when making assessments. Though only two-layer quality scalability was subjectively tested, proposals supporting finer granularity scalability were encouraged, and were evaluated by expert panels.

Table 4 gives, for each test, the form of scalability (object, temporal quality, or spatial quality), the test sequence class and test sequences, the test bitrate(s), the test method, the decoded sequence display format, the reference sequence format, the coding technique used for the anchor sequence, the number of proposals received to the test, and the number of test sessions used for the test. Table 4 also indicates for each test sequence, the "designated object(s)" to which scalability was applied. A proposer could address any subset of the tests listed in Table 4. However, for each test addressed, a proposer was required to encode all the test sequences at all the bitrates for that test, using only the specified form of scalability (one row in Table 4). Anchors were not used in the quality scalability tests, since no standard technique having this functionality was readily available. For tests using a double stimulus method, the reference sequence was the original uncoded test sequence.

| Sequence class: | Test sequences (designated objects) | Functionality | Bit rates | Test method | Display format | Reference format | Anchor coding technique | Num. of proposals/ Num. of test sessions |
|---|---|---|---|---|---|---|---|---|
| A: | Akiyo (woman) Sean (man), Hall monitor (man with monitor) | Object scalability | 48 | SS | CIF(1) | N/A | H.263(1) | 10/1 |
| B/C: | Stefan (tennis player) News (dancers in the monitor), Coast guard (big boat) | Object scalability | 1024 | DSCQS | ITU-R 601 | ITU-R 601 | MPEG1(2) | 3/1 |
| E: | Weather (woman), Children (children), Bream (fish) | Object scalability | 320 | DSIS | ITU-R 601 | ITU-R 601 | MPEG1(2) | 2/1 |
| A: | Akiyo (woman), Sean (man), Hall monitor (man with monitor) | Spatial quality scalability | 24, 24+24 | SS | CIF(1) | N/A | N/A | 2/2 |
| | Stefan (tennis player), Children | | | | | | | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| B/C/E: | (children), News (dancers in the monitor), Coast guard (big boat) | Spatial quality scalability | 512, 512+512 | DSCQS | ITU-R 601 | ITU-R 601 | N/A | 3/2 |
| A: | Akiyo (woman), Sean (man), Hall monitor (man with monitor) | Temporal quality scalability | 24, 24+24 | SS | CIF[1] | N/A | N/A | 1/2 |
| B/C/E: | Stefan (tennis player), Children (children), News (dancers in the monitor), Coast guard (big boat) | Temporal quality scalability | 512, 512+512 | DSCQS | ITU-R 601 | ITU-R 601 | N/A | 1/2 |

[1] CIF displayed in a window centered in the 60 Hz ITU-R 601 format.

[2] CIF upsampled to 60 Hz ITU-R format.

Table 4 - Content-based scalability tests.

For the quality scalability tests, proposers were required to provide two versions of each test sequence: (1) For the base layer tests, proposers provided each test sequence coded at the base layer bitrate. For example, for a proposal to the last test listed in Table 4 ("sequence class B/C/E, temporal quality scalability"), a decoded version of each test sequence at 512 kbit/s was required. There were no restrictions placed on the spatio-temporal resolutions chosen for the objects in the base layer sequence (each object could have a different spatio-temporal resolution). (2) For the base+enhancement layer tests, proposers provided each test sequence coded at the base+enhancement layer bitrate. The additional (beyond the base layer) bitrate was required to be applied to enhance the quality of the designated object(s). This enhancement could be achieved only via the form of scalability specified in Table 4 (either spatial or temporal resolution changes, but not both), and was required to be accomplished in a manner independent of the coding of rest of the video scene. Proposers were required to clearly describe in the proposal technical description the mechanisms used to enhance the designated object(s). Continuing the example above, the base+enhancement layer test required a decoded version of each test sequence at 1024 kbit/s, where the additional 512 kbit/s was used to enhance the temporal resolution of the designated object(s) in each sequence.

To extract the designated object(s) from the test sequences, as mentioned in Section 4, proposers were encouraged to use the segmented test sequences provided as part of the test materials. Other segmentations were permitted, provided that the segmentation technique was clearly described in the proposal technical description

# 8. Statistical Analysis

The primary goal of the first round of the MPEG-4 video subjective tests was to accurately rank the proposals to each test according to their performance. In order to obtain these rankings, and to fully understand the reliability of the rankings, two classes of statistical analyses were performed on the test data.

First, for each codec (proposal or achor) in each test, elementary statistics (mean, standard deviation, confidence interval) were computed over all test subjects for each test sequence, and also over all test subjects and test sequences. The codecs were ranked based on their mean scores, and a complete Student analysis was used to determine which codecs' mean scores were statistically significantly different. Section 8.1 details the statistical analyses performed to rank the codecs, and describes the tables and graphs used to illustrate the results. A second set of statistical analyses validated the test data, and further investigated other sources of variation in the test data. Section 8.2 describes this analysis and summarizes its outcomes.

Since presentation of the results of the evaluations of the MPEG-4 proposals is outside the scope of this paper, the actual results of the MPEG-4 subjective test data analysis are not reported here. The complete results of the MPEG-4 subjective tests are contained in "Report of the ad hoc group on MPEG-4 video testing logistics" [4].

8.1 Codec Ranking

As described in Section 2, test subjects expressed their opinions of the codecs via ratings on several different types of scales. The first step in the statistical analysis of the test data was to convert these ratings to numerical values as follows:

DSCQS - For each sequence in a trial (i.e. for the reference sequence and the processed sequence), the position of the test subject's mark on the rating scale (see Figure 2) was linearly converted to an integer value ranging from 0 (corresponding to the bottom of the rating scale) to 100 (corresponding to the top of the rating scale). The score for the trial was then computed as the difference between the reference sequence rating and the processed sequence rating. Thus, a score close to zero indicates an algorithm with good performance (having quality close to that of the reference, original sequence). Note that with this scoring technique, negative scores are possible, when a test subject attributes higher quality to the coded sequence than to the reference.

DSIS - The five rating levels (see Fig. 4) were converted to integer values ranging from 1 (corresponding to the level labeled "Very Annoying") to 5 (corresponding to the level labeled "Imperceptible").

DSBV - Check marks in "YES" boxes were converted to 1, and in "NO" boxes to 0.

SS - The eleven rating levels (see Fig. 6) were converted to integer values ranging from 0 (corresponding to the level below "Bad") to 10 (corresponding to the level above "Excellent").

After ratings conversions, statistical analyses were performed on the data from each test separately. For each bitrate/error condition, for each codec, for each sequence, the mean score, the standard deviation, and the 95% confidence interval were calculated, over all test subjects. These statistics were also calculated over all test subjects and test sequences. Finally, the mean scores for each codec were averaged over all bitrates/error conditions. Codecs were ranked according to their mean score over all sequences and subjects.

In order to determine when the mean scores of two codecs were statistically significantly different (SSD), a complete Student analysis was performed on the data for each test, for each bitrate/error condition. This analysis provided a probability of equivalence of the mean scores for each pair of codecs. Codecs having mean scores with a probability of equivalence lower than 0.05 were judged to be SSD. When two codecs are not SSD, even if they have different mean scores, nothing can be concluded about their relative performance in the tests. There is not sufficient precision in the statistical estimate of their mean scores to allow relative ranking of the two.

The statistical results were presented via a set of tables like that shown in Table 5. Each row in this table presents the data for one tested codec (proposal or anchor). Each column (except the rightmost two) presents the statistics for one test sequence: mean, standard deviation, and confidence interval. The column labeled "mean" gives a codec's statistics over all test sequences, and the rightmost column identifies the next codec in the table with an SSD mean score. Figure 10 shows a bar chart graphically presenting the means and confidence intervals of the codecs listed in Table 5. In the example table and graph shown here, names of the proposers and the anchor have been replaced with letters.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | | | |
| | | | | | | | Next statist. |
| | | Foreman | News | Silent_V | Coast_G | Mean | different |
| | Mean | 6,67 | 8,07 | 7,97 | 6,03 | 7,18 | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| A | St.Dev. | 2,01 | 1,39 | 1,4 | 2,08 | 1,93 | F |
| | C.I. | 1,02 | 0,7 | 0,71 | 1,05 | 0,49 | |
| | Mean | 7 | 8,13 | 8 | 4,33 | 6,87 | |
| B | St.Dev. | 1,78 | 1,46 | 1,44 | 1,88 | 2,24 | G |
| | C.I. | 0,9 | 0,74 | 0,73 | 0,95 | 0,57 | |
| | Mean | 5,6 | 7,5 | 7,27 | 6,23 | 6,65 | |
| C | St.Dev. | 2,08 | 1,76 | 2,05 | 2,1 | 2,12 | G |
| | C.I. | 1,05 | 0,89 | 1,04 | 1,06 | 0,54 | |
| | Mean | 7,5 | 6,97 | 7,03 | 5,03 | 6,63 | |
| D | St.Dev. | 1,8 | 1,94 | 1,35 | 1,65 | 1,93 | G |
| | C.I. | 0,91 | 0,98 | 0,68 | 0,84 | 0,49 | |
| | Mean | 5,03 | 7,87 | 6,77 | 6,33 | 6,5 | |
| E | St.Dev. | 2,11 | 2,06 | 1,89 | 2,35 | 2,32 | I |
| | C.I. | 1,07 | 1,04 | 0,96 | 1,19 | 0,59 | |
| | Mean | 5,7 | 7,03 | 6,53 | 5,57 | 6,21 | |
| F | St.Dev. | 1,97 | 1,94 | 1,43 | 2,1 | 1,95 | I |
| | C.I. | 1 | 0,98 | 0,72 | 1,06 | 0,49 | |
| | Mean | 4,73 | 6,67 | 6,53 | 5,6 | 5,88 | |
| G | St.Dev. | 2,15 | 2,07 | 2,03 | 1,98 | 2,18 | I |
| | C.I. | 1,09 | 1,05 | 1,03 | 1 | 0,55 | |
| | Mean | 5,63 | 5,93 | 6,07 | 5,77 | 5,85 | |
| H | St.Dev. | 1,9 | 2,02 | 1,78 | 2,06 | 1,93 | I |
| | C.I. | 0,96 | 1,02 | 0,9 | 1,04 | 0,49 | |
| | Mean | 4 | 5,17 | 5,57 | 4,87 | 4,9 | |
| I | St.Dev. | 1,78 | 2,2 | 2,24 | 2,01 | 2,12 | J |
| | C.I. | 0,9 | 1,11 | 1,13 | 1,02 | 0,54 | |
| | Mean | 3,3 | 3,17 | 3,3 | 3,5 | 3,32 | |
| J | St.Dev. | 1,74 | 1,72 | 1,78 | 1,57 | 1,69 | M |
| | C.I. | 0,88 | 0,87 | 0,9 | 0,79 | 0,43 | |
| | Mean | 2,67 | 4,1 | 4,17 | 2,23 | 3,29 | |
| K | St.Dev. | 1,06 | 1,54 | 1,39 | 0,82 | 1,49 | M |
| | C.I. | 0,54 | 0,78 | 0,7 | 0,41 | 0,38 | |
| | Mean | 1 | 3,87 | 5,07 | 1,83 | 2,94 | |
| L | St.Dev. | 1,08 | 1,63 | 2,03 | 1,26 | 2,22 | M |
| | C.I. | 0,55 | 0,83 | 1,03 | 0,64 | 0,56 | |
| | Mean | 2,57 | 2,07 | 2,53 | 1,23 | 2,1 | |
| M | St.Dev. | 1,48 | 1,39 | 1,55 | 1,07 | 1,47 | |
| | C.I. | 0,75 | 0,7 | 0,78 | 0,54 | 0,37 | |

TABLE 5 - Example of table with SS method results

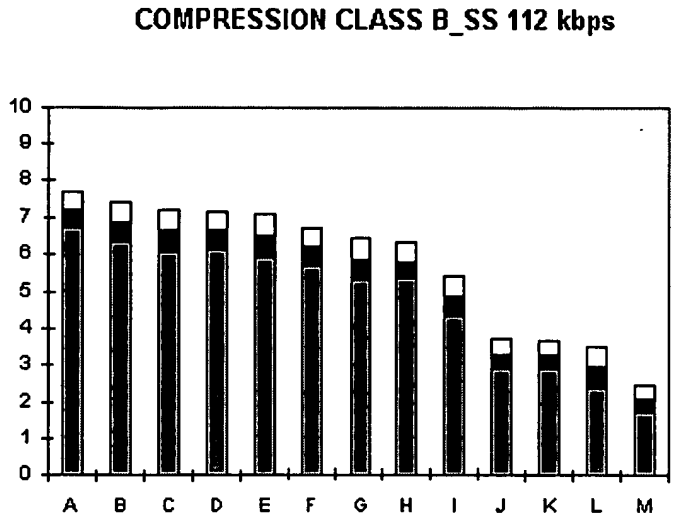



Figure 10 - Example of bar chart with SS method results

8.2 Test Data Validation

During the statistical processing of the MPEG-4 subjective test data, a rejection criterion of inconsistent test subjects was applied, according to [3]. No test subject was rejected, showing that the test subjects' behaviors were reliably stable, within each test session, and over all sessions of each test.

Statistical analyses were performed to investigate the distribution of the data for each MPEG-4 test. The Beta 2 test [3] verified a close-to-Gaussian distribution on the test data. The generally centered nature of the test data distributions indicated that at least the parts of the rating scales used in the tests were well suited to the tests. However, the test data distributions also revealed problems with some test methods. In some DSCQS tests, the test data distribution did not completely cover the available scoring range, even though it was well centered in the score range. This is attributable to large differences in quality between the test and reference sequences. The distribution of the DSBV test data indicated that the performances of the proposals in those tests were not sufficiently distinguishable. The test results were characterized by a very high overall mean opinion score and a high overall standard deviation. The high mean opinion score was attributable to the fact that most of the codecs were able to recover from the channel errors, while the high standard deviation can be explained by the binary scoring method.

An analysis of variance (ANOVA) was performed on the data from each test session. This analysis checked for any codec, test sequence, or test subject dependence in the test data. The results of the ANOVA showed that, for test sessions of similar size, the sequence and observer variances within the session were reasonably low and stable. ANOVA also verified that the codec effect primarily explained the overall variance in the codecs' mean scores. Variance due to test subject and test sequence (separately) were also significant, though at a lower level than the codec effect. The variance due to test subject was almost certainly because of the use of video expert test subjects, whose assessments of quality are typically found to be acceptably consistent across sessions, but noticeably different from one subject to another. This may be explained by the ability of experts to recognize different coding techniques by their characteristic artifacts.

In order to further investigate the influence of test sequence on the test results, within each test, the discriminating power for each sequence was estimated from a Student analysis of the standard deviation of each sequence over the different codecs. When a sequence is more discriminating, the difference in its scores between codecs is higher, thus its standard deviation is also higher. This analysis showed that, for class A sequences "Akiyo" was the most discriminating, and "Mother & daughter" was the least discriminating. For class B sequences, "News" was the most discriminating, and for class C it was systematically the sequence "Stefan". Within each test, using a Student analysis of the mean scores over

the codecs for each sequence, the criticality of the sequence was also analyzed. A more critical sequence is one that proves more difficult to code. The average score over all codecs of such a sequence is significantly worse than that of the other sequences. No MPEG-4 test sequence was found to be significantly more critical than others.

Comparing the grand means for each session, no particular trend was detected. This clearly shows two things. First, it shows that, even for the tests consisting of several separate test sessions, no fatigue and no learning effects were detectable in the test data. It also shows that random ordering of the trials used to build each test session was effective. There was also very good coherence among the test sessions in terms of spread of video quality, indicating no contextual effects in the data. This indicates that the use of two presentation orders for each SS test was effective.

# 9. Evaluations and Conclusions

The tests described in this paper were designed to compare the MPEG-4 proposers' codecs to one another and to existing standard codecs, in terms of subjective quality. Many factors (functionality, bitrate, complexity of test sequence, etc.) had to be considered in order to specify fair and effective subjective tests. Standard subjective test methods were tailored to the goals of each test, or in some cases new test methods were even designed. The test methods described here generally performed well, and the reliability of the test results was generally high. The tests provided the MPEG video experts with relevant information for use in developing the first version the MPEG-4 Video Verification Model. The success of these tests is especially significant, since they were the first major subjective tests conducted at these bitrates and for these functionalities. Besides obtaining information for use in development of the MPEG-4 standard, the test data and the comments from the test subjects also yielded other valuable insights, about which tests were more or less effective in achieving their goals, and about how the test methods could be improved upon for future use.

Concerning the DSCQS test method, as mentioned in Section 8.2, the difference in quality between the proposers' sequences and the reference sequences was probably too high in some cases, reducing the effectiveness of the reference sequences. There are two possible solutions to this problem. The first is to use an impaired reference instead of the original CCIR 601 resolution sequence. Appropriate impairment(s) to produce useful references must be studied and validated. The second solution is to simply use the SS test method for all bitrates, even the relatively high ones.

In the error recovery tests, nearly all proposers' decoded sequences recovered from the burst error. Since proposers knew beforehand how the evaluation would be performed, a strategy was relatively easy to devise for recovering from the error burst (for example, simply insert an intra-coded frame towards the end of the sequence). Even though recovery strategy has a noticeable impact on the subjective quality of a decoded sequence, the test did not assess this, which made it difficult to fully evaluate the performance of the proposed codecs. For these tests, it would be more effective to use an evaluation where decoded sequence quality is continuously monitored over time, in order to get a subjective assessment of the overall quality of the sequences. A recommendation on this type of test has been recently introduced in the new draft ITU-R Recommendation BT.500-7 [12]. MPEG is currently working to adapt these methods to the specific MPEG-4 test requirements.

Concerning the content-based quality scalability tests, it is not clear that the designated object(s) quality assessments were entirely effective. In assessing the quality of the designated object(s), the judgment of a test subject is likely to be undesirably influenced by the background surrounding the object(s). Furthermore, in assessing global quality, the evaluation criteria could vary between subjects, depending on how each subject's attention is divided between background and designated object(s). Presentation of just the designated object(s) on a uniform gray background would avoid these effects.

For most of the MPEG-4 subjective tests, proposers had a great deal of freedom of choice about such things as pre- and post-processing techniques, down- and up-sampling filters, and spatial and temporal coding resolutions. However, the resulting differences among proposals, especially with regards to temporal resolution, created some difficulties in proposal comparisons. Proposals that traded-off low temporal resolution for improved spatial quality tended to receive better subjective ratings than proposals that used higher temporal resolution at the expense of spatial quality. Test subjects seemed to not always consider temporal resolution when making assessments. This could indicate a basic fact about the relative subjective importance of temporal and spatial resolution at low bitrates. However, it may also indicate that a set of test sequences with larger amplitude motion should be included in future tests.

Concluding, we note that there will be a second round of MPEG-4 video subjective tests in July '97. In these tests, new proposals will be compared to one another, as well as to the MPEG-4 Video Verification Model, and to already existing standards. In preparation for these new tests, building on the useful experience gained from the tests described in this paper, MPEG-4 is refining the test methods described here, and defining new test methods. The work is now concentrated on the evaluation of such things as video quality in the presence of audio, error robustness in long sequences, and object-based quality, all of which are important to MPEG-4, and to future coding schemes in general.

# Acknowledgments

T. Alpert, L.Contin, R. Koenen and F. Pereira acknowledge the support of the European Commission under the RACE program - project MAVT R2072, and the ACTS program - project TAPESTRIES AC055. F. Pereira acknowledges the support of Junta Nacional de Investigação Científica e Tecnológica for the support of this work under the project 'Processamento Digital de Áudio e Vídeo'.

We also acknowledge the many people who contributed to the success of these tests. We are particularly grateful to colleagues that participated as test administrators or as test subjects.

# References

[1] J. Ostermann (editor), Report on the ad hoc group on the evaluation of tools for non tested functionalities of video submissions, Doc. ISO/IEC JTC1/SC29/WG11 N1064 Dallas meeting, November 1995

[2] J. Ostermann, Methodology used for the evaluation of video tools and algorithms in MPEG-4, Image Communication, this issue.

[3] ITU-R, Methodology for the subjective assessment of the quality of television pictures Recommendation BT.500-6, 1994

[4] H. Peterson (editor), "Report of the Ad Hoc Group on MPEG-4 Video Testing Logistics", Doc. ISO/IEC JTC1/SC29/WG11 N1056, Dallas MPEG meeting, November 1995

[5] ITU-R, "Specifications and Alignment Procedures for Setting of Brightness and Contrast of Displays", Recommendation BT. 814-1 - 1994

[6] ITU-R, "Encoding parameters of digital television for studios", Recommendation BT. 601-4, 1994

[7] MPEG-4 Video Verification Model - Version 1.0, ISO/IEC JTC1/SC29/WG11 N1172, Munich MPEG meeting, January 1996.

[8] F. Pereira (editor), MPEG4 testing and evaluation procedures document, Doc. ISO/IEC JTC1/SC29/WG11 N999, Tokyo MPEG meeting, July 1995

[9] ITU-T, Video coding for low bitrate communication, Recommendation H.263, 1996

[10] ISO/IEC IS 11172-2 (MPEG-1 Video), Coding of moving pictures and associated audio for digital storage media up to about 1.5 Mbit/s, 1993

[11] MPEG AOE Group, Proposal Package Description (PPD) - Revision 3, Doc. ISO/IEC JTC1/SC29/WG11 N998, Tokyo meeting, July 1995

[12] ITU-R, Methodology for the subjective assessment of the quality of television pictures, Recommendation BT.500-7, 1995